

Maximum-A-Posteriori Estimates in Linear Inverse Problems with Log-concave Priors are Proper Bayes Estimators

Martin Burger^{1,2}, Felix Lucka^{1,2,3}

¹ Institute for Computational and Applied Mathematics, University of Münster, Einsteinstr. 62, D-48149 Münster, Germany

² Cells in Motion Cluster of Excellence, University of Münster, Mendelstr. 12, D-48149 Münster, Germany

³ Institute for Biomagnetism and Biosignalanalysis, University of Münster, Malmedyweg 15, D-48149 Münster, Germany

E-mail: martin.burger@wwu.de

Abstract. A frequent matter of debate in Bayesian inversion is the question, which of the two principle point-estimators, the maximum-a-posteriori (MAP) or the conditional mean (CM) estimate is to be preferred. As the MAP estimate corresponds to the solution given by variational regularization techniques, this is also a constant matter of debate between the two research areas. Following a theoretical argument - the Bayes cost formalism - the CM estimate is classically preferred for being the Bayes estimator for the mean squared error cost while the MAP estimate is classically discredited for being only asymptotically the Bayes estimator for the uniform cost function.

In this article we present recent theoretical and computational observations that challenge this point of view, in particular for high-dimensional sparsity-promoting Bayesian inversion. Using Bregman distances, we present new, proper convex Bayes cost functions for which the MAP estimator is the Bayes estimator. We complement this finding by results that correct further common misconceptions about MAP estimates. In total, we aim to rehabilitate MAP estimates in linear inverse problems with log-concave priors as proper Bayes estimators.

AMS classification scheme numbers: 65J22, 62F15, 62C10, 65C60, 62F10, 65C05

Submitted to: *Inverse Problems*

1. Introduction

Bayesian models have received considerable attention in inverse problems over the last years. A particular advantage of the Bayesian approach is the systematic treatment of stochastic forward models and of prior knowledge about solutions, which is closely related to classical regularization theory. While the basic idea is now widely accepted in the inverse problems community, there is still a debate concerning the choice of point estimates. While pragmatical and computational reasons are clearly favouring maximum a-posteriori probability estimates, those are considered inferior to others like conditional mean estimates by statistical arguments. In particular the Bayes cost approach argues the latter to minimize a natural cost while the maximum a-posteriori probability estimate can only be obtained asymptotically from a degenerate cost. In this paper, we will present a novel viewpoint on maximum a-posteriori probability estimates, having in mind high-dimensional log-concave priors such as popular sparsity priors. Our computational and theoretical results puts the inferiority compared to conditional means estimates under question.

We consider the inverse problem of solving a linear, ill-posed operator equation for the true, infinite-dimensional solution \tilde{u} . Here, we start from the following discrete model chosen for obtaining a computational solution:

$$f = K u + \varepsilon, \quad (1)$$

where $f \in \mathbb{R}^m$ represents the given measured data, $u \in \mathbb{R}^n$ represents a discretization of \tilde{u} , $K \in \mathbb{R}^{m \times n}$ is the discretization of the continuous forward operator with respect to the spaces of u and f and $\varepsilon \in \mathbb{R}^m$ is an additive, stochastic noise term. For simplicity we restrict ourselves to the case of ε being Gaussian. We want to solve (1) in the framework of Bayesian inversion, which we will briefly sketch in the following (cf. [21] for further details and [15, 6, 20, 16, 25, 37] for exemplary applications):

First, the stochastic nature of the noise term renders (1) into a relation between the random variables F and \mathcal{E} :

$$F = K u + \mathcal{E}, \quad (2)$$

where we assume $\mathcal{E} \sim \mathcal{N}(0, \Sigma_\varepsilon)$. Now, (2) determines the conditional probability density of F given u (the *likelihood* density):

$$p_{li}(f|u) \propto \exp\left(-\frac{1}{2}\|f - K u\|_{\Sigma_\varepsilon^{-1}}^2\right), \quad \text{with} \quad \|y\|_A^2 := y^T A y \quad (3)$$

Standard statistical inference strategies like *maximum-likelihood estimation* would try to estimate u on the basis of (3). However, in typical inverse problems, the ill-posedness of (1) precludes this approach. *Bayesian inference strategies* rely on encoding *a-priori* information about u by modeling it as a random variable as well (U in our notation). Its density, $p_{pr}(u)$ is therefore called the *prior*. *Bayes' rule* can then be used to construct the posterior probability density:

$$p_{po}(u|f) = \frac{p_{li}(f|u)p_{pr}(u)}{p(f)} \quad (4)$$

This conditional density of U given F is called the *posterior*. In Bayesian inversion, this density is the complete solution to the inverse problem and, thus, the central object of interest (see Figure 1 for an illustration). *Bayesian inference* is the process of extracting the information of interest from the posterior:

- Point estimates infer a single estimate of u from the posterior.

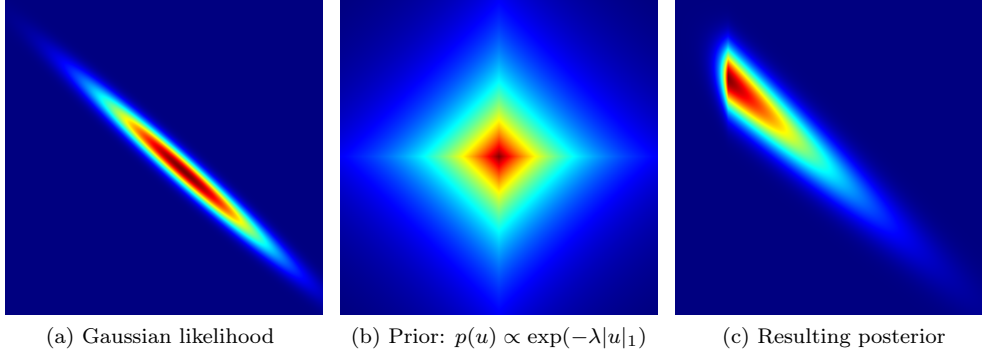


Figure 1: Illustration of possible shapes of likelihood, prior and posterior for $n = 2$.

- Credible regions estimates search for sets that bound u with a certain probability.
- Extreme value probabilities try to estimate the probability that a feature $g(u)$ exceeds some critical value.
- Conditional covariance estimates try to assess the spatial distribution of variance and dependencies between the components of u .
- Histogram estimates analyze the distributions of single components u_i .

More advanced Bayesian techniques like the treatment of nuisance parameters by *marginalization* or *approximation error modeling* [21, 22, 33, 28, 37], *model comparison, selection or averaging* [39, 19], and *experimental design* [38] are also based on the above formalism and principles.

Up to now, we did not address the most important step in Bayesian inversion, i.e., the construction of the prior $p_{pr}(u)$ (*Bayesian modeling*). All theorems developed in this work are valid for *log-concave Gibbs distributions* of the form

$$p_{pr}(u) \propto \exp(-\lambda \mathcal{J}(u)), \quad (5)$$

where $\mathcal{J}(u)$ is a convex functional (called the prior *energy*), and $\lambda > 0$ is a scaling parameter. This includes a wide range of distributions commonly used in Bayesian inversion. The corresponding posterior is given by:

$$p_{po}(u|f) \propto \exp\left(-\frac{1}{2}\|f - Ku\|_{\Sigma_\varepsilon^{-1}}^2 - \lambda \mathcal{J}(u)\right) \quad (6)$$

We will need some further, but rather technical properties later, which are fulfilled for all commonly used convex $\mathcal{J}(u)$.

1.1. MAP vs. CM Estimates: Variational Regularization vs. Bayesian Inference?

Choosing a single point estimate for u is the most simple but also most commonly used Bayesian inference technique. Two popular estimates are:

- Maximum a-posteriori-estimate (MAP):

$$\hat{u}_{\text{MAP}} := \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \{ p_{po}(u|f) \} \quad (7)$$

We can compute \hat{u}_{MAP} for (6) by solving a high-dimensional optimization problem:

$$\hat{u}_{\text{MAP}} = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2} \|f - K u\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda \mathcal{J}(u) \right\} \quad (8)$$

This is a *Tikhonov-type* regularization of equation (1) [9]. Hence, MAP estimation yields a direct correspondence to *variational regularization* techniques [3, 35].

- Conditional mean-estimate (CM):

$$\hat{u}_{\text{CM}} := \mathbb{E}[u|f] = \int u p_{po}(u|f) du \quad (9)$$

Computing \hat{u}_{CM} requires solving a high-dimensional *integration* problem [21, 36].

The immediate and obvious question is: What is the difference between MAP and CM estimate? Which of them is "better" in general, or for a specific task? This is not only a matter of constant debate within the field of Bayesian inversion, but also with classical regularization theory due to the direct correspondence of \hat{u}_{MAP} . This article starts with a summary of the "classical" view on the issue. Then, several recent computational and theoretical results are discussed, which challenge this point of view. In the last part, new theoretical ideas are introduced that fit to all of these results, disprove certain common myths and will lead to new insights and perspectives for the comparison of variational regularization and Bayesian inference.

1.2. Sparsity Constraints in Inverse Problems

Sparsity constraints are a type of a-priori information that demand the solution of (1) to have very few non-zero coefficients in a suitable representation (i.e., a bases, frames or other dictionaries). Solving high-dimensional inverse problems using sparsity constraints has led to enormous advances in various areas, a popular example being *total variation* (TV) deblurring [4], based on sparsity constraints on the gradient of the unknown quantity. Commonly, sparsity constraints are formulated in the framework of variational regularization by choosing ℓ_1 -type norms for $\mathcal{J}(u)$ in (8):

$$\hat{u}_\alpha = \operatorname{argmin}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2} \|f - K u\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda |Du|_1 \right\} \quad (10)$$

Recently, using similar sparsity-constraints in Bayesian inversion has attracted considerable attention. There are two common ways to encode sparsity in the prior:

- (i) Converting the functionals used in (10) directly into priors of the form $p_{pr}(u) \propto \exp(-\lambda |Du|_1)$ (ℓ_1 -type priors). This is convenient, since the prior is log-concave and one already knows that the MAP estimate will be sparse. In this article, we will only present computational results for this type of priors (the theoretical results are, however, valid for all log-concave priors). Note that such priors are not a multivariate generalization of Laplace distributions (see, e.g., [8]).
- (ii) Hierarchical Bayesian modeling (HBM) extends the prior model by an additional level and imposes sparsity on this level. While this approach has shown to yield good results in various applications (e.g., [1, 31, 34, 40]), a potential difficulty is that the implicit priors over the unknowns are usually not log-concave.

Figure 2 shows random images drawn from a Gaussian, a ℓ_1 -type and a Student's t-distribution (a potential implicit prior encountered in HBM). While the visual impression of the ℓ_1 random image clearly differs from the Gaussian one, it is by

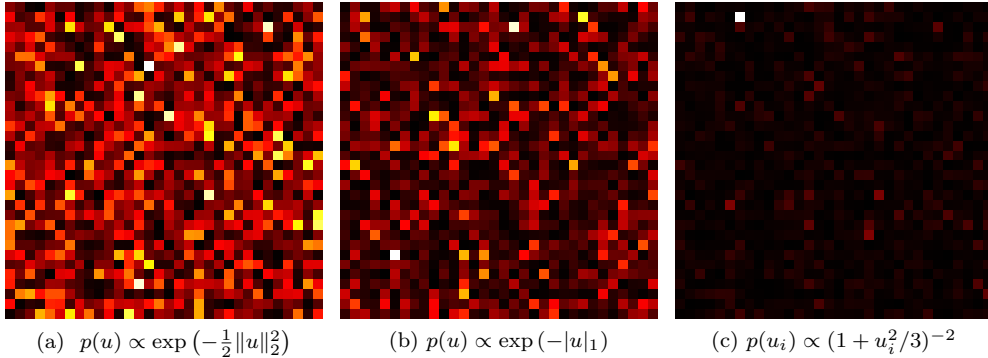


Figure 2: Random images draws from different prior distributions. Here, u was assumed to correspond to a 2D image of $n = 32 \times 32$ pixels $u_1, \dots, u_{(32^2)}$.

no means sparse in the traditional sense. If it would be the true solution, \hat{u}_{MAP} would probably not be able to recover it in a satisfactory way. This is due to misconceptions behind the "reverse engineering" approach of turning a sparsifying regularization functional $\mathcal{J}(u)$ into a prior $p_{pr} \propto \exp(-\lambda\mathcal{J}(u))$. This has already been noticed in [12, 13]. More general, it points to an inherent difficulty of defining sparsity in the Bayesian framework in a meaningful, consistent and tractable way, which we cannot address further in this article.

The paper is structured as follows: In Section 2, we will provide a discussion of the comparison between MAP and CM estimates. First, we will revisit the classical view on the problem in Section 2.1, which favors the CM estimate on the basis of a theoretical argument: The Bayes cost formalism. While the CM estimate minimizes the mean squared error as a cost function, the MAP estimate is discredited for minimizing a binary cost function in its degenerate limit. Then, in Section 2.2, we summarize recent results and observations, which do not fit in this picture. In particular, these results steam from linear inverse problems with sparsity-promoting priors incorporating ℓ_1 -type norms and motivate the further theoretical investigation in Section 3. There, we will present new, proper, convex Bayes cost functions for both MAP and CM estimates. The key ingredient will be the use of *Bregman distances*. This renders the MAP estimate into a proper Bayes estimator and fits to the otherwise contradictory computational observations. In addition, we present further results that correct common misconceptions about MAP estimates.

2. MAP vs. CM Estimates

In the following we start by briefly discussing the established viewpoint on the comparison of MAP and CM estimates and subsequently review several recent results converse to this viewpoint, supplemented by some computational experiments. The CM estimate is the mean of the posterior, while the MAP estimate is the (highest) mode of the posterior (see Figure 3). However, this does not provide any intuition why one of them should be better suited to represent a distribution. Hence, a lot of presentations of the topic provide plots of hypothetical distributions like Figure 4 to show that none of them is better in general. However, one might argue that the CM

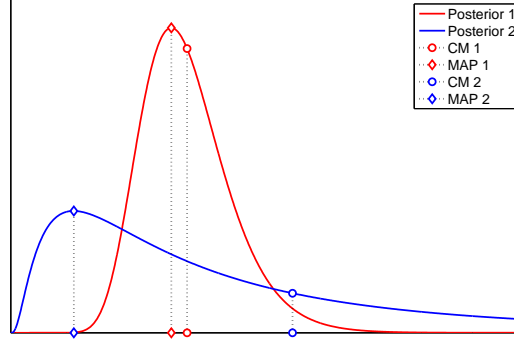


Figure 3: Comparison of MAP and CM estimates for two posterior densities.

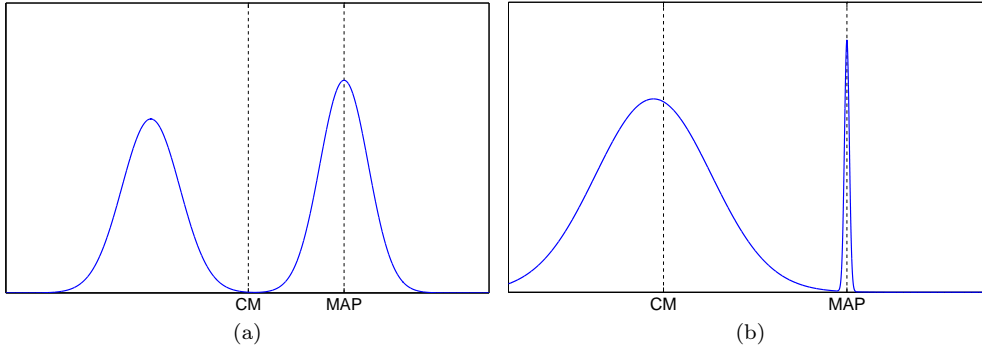


Figure 4: Hypothetical, bimodal distributions to show that none of the estimates is better in general.

estimator as the mean value is an intuitive choice as it is the "center of (probability) mass" and corresponds to the average of a sample, familiar from every-day descriptive statistics.

2.1. Bayes Cost Formalism

As the illustrative comparison does not give any useful intuition, the Bayes cost formalism is usually used to provide a decisive theoretical argument, which we recall in the following (cf. [21, 23]). In the Bayesian framework, an estimator \hat{U} is a random variable as well, as it relies on F and U . *Statistical estimation theory* (respectively Bayesian decision theory) examines the general behavior of estimators to find optimal estimators for a given task. A common approach is to define a *cost function* $\Psi(u, \hat{u})$ measuring the desired and undesired properties of \hat{u} . The *Bayes cost* is defined by the expected cost, i.e., the average performance:

$$\begin{aligned} BC_{\Psi}(\hat{u}) &:= \mathbb{E}[\Psi(u, \hat{u}(f))] = \int \int \Psi(u, \hat{u}(f)) p(u, f) \, du \, df \\ &= \int \int \Psi(u, \hat{u}(f)) p_{\text{like}}(f|u) \, df \, p_{\text{prior}}(u) \, du \end{aligned}$$

$$\stackrel{(4)}{=} \int \int \Psi(u, \hat{u}(f)) p_{po}(u|f) du p(f) df \quad (11)$$

The Bayes estimator \hat{u}_Ψ is the estimator, which minimizes $BC_\Psi(\hat{u})$.

$$\hat{u}_\Psi := \underset{\hat{u}}{\operatorname{argmin}} \{BC_\Psi(\hat{u})\}$$

In (11), $\hat{u}(f)$ only depends on f and the marginal density $p(f)$ is non-negative. Thus, \hat{u}_Ψ also minimizes

$$\hat{u}_\Psi(f) = \underset{\hat{u}}{\operatorname{argmin}} \left\{ \int \Psi(u, \hat{u}(f)) p_{po}(u|f) du \right\} \quad (12)$$

The main classical arguments in favour of CM and against MAP estimates derived from the Bayes cost formalism are as follows:

- The CM estimate is Bayes estimator for the mean squared error

$$\Psi_{\text{MSE}}(u, \hat{u}) = \|u - \hat{u}\|_2^2, \quad (13)$$

which seems to be a very natural and reasonable choice for Ψ . Interpreted geometrically, one also speaks of a "well-centeredness" of $p_{po}(u|f)$ around \hat{u}_{CM} . As it is by default unbiased with respect to $p_{po}(u|f)$, one can further show that \hat{u}_{CM} is also the *minimum error variance estimator*.

- On the other hand, the MAP estimate can only be seen as an *asymptotic* Bayes estimator of

$$\Psi_\delta(u, \hat{u}) = \begin{cases} 0, & \text{if } \|u - \hat{u}\|_\infty \leq \delta \\ 1 & \text{otherwise} \end{cases} \quad (14)$$

for $\delta \rightarrow 0$ (*uniform cost* or *0-1 loss*). Thus, it is usually not considered a proper Bayes estimator. This characterization also does not seem to allow for an intuitive geometrical interpretation of \hat{u}_{MAP} akin to the one for \hat{u}_{CM} .

The theoretical difference between MAP and CM estimates seems to fit to their different computational complexity (cf. Section 1.1). The theoretical discrimination of the MAP estimate contrasts with its practical usability, in particular for high-dimensional inverse problems. In Bayesian inversion for such one therefore often encounters a strange contrariness: Usually a careful prior modeling is introduced, for which the CM estimate is regarded as optimal. However, for computational reasons, a MAP instead of a CM estimate is computed, which is rather regretted and excused for. If the results turn out to be not fully satisfactory, the discussion identifies shortcomings of the MAP estimate to be responsible for this. Even if the results turn out to be really good, concern that using MAP estimates is not a proper Bayesian technique is often expressed.

2.2. Converse Results

In the following we discuss some recent theoretical results and computational experiments indicating that CM estimates are not superior in particular for high-dimensional inversion. We start with the Gaussian case: Gaussian priors are the most popular and arguably the most fundamental class of priors one can consider for (6), due to various reasons such as their maximum entropy property, alpha-stability and the central limit theorem. However, for this most fundamental class of priors, the seemingly fundamentally different MAP and CM estimate happen to be equal.

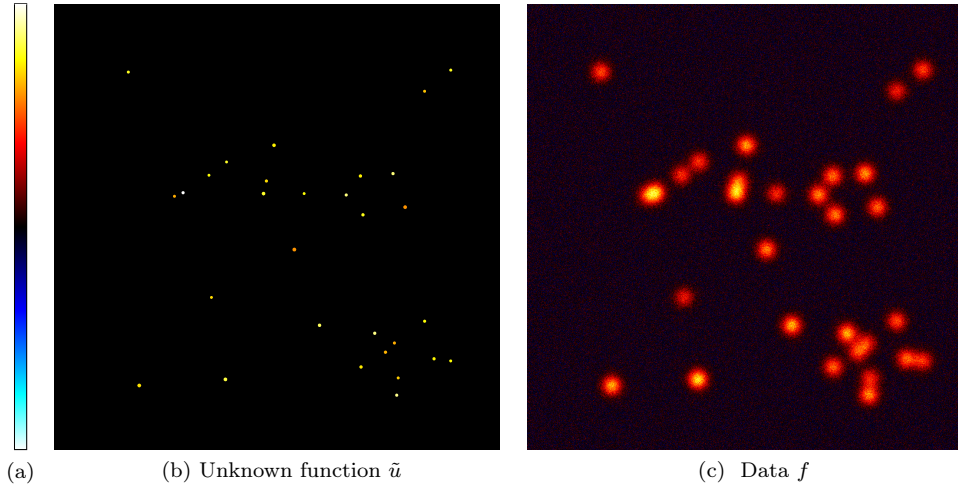


Figure 5: A simple 2D deblurring example.

From the classical view, this can only be interpreted as a funny coincidence, which is arguably not fully satisfactory. One might argue however that a quadratic Bayes cost is perfectly suited for Gaussian models, since it corresponds well to the negative logarithm of the prior. An appropriately scaled square-norm Bayes cost as used above should obviously be quite robust also for high-dimensional problems in the case of Gaussian priors, while this is not clear at all for other priors. For nonquadratic priors asymptotically concentrating on some Banach space it is not at all clear whether there is a robust and asymptotically meaningful squared norm in the Bayes cost criterion, hence it might be reasonable to think about other costs better suited to the Banach space limit. We will further dwell upon this issue in the Section 3, and continue with results from computational experiments with ℓ_1 type priors (cf. Section 1.2). We will use the Split Bregman method [11] for computing MAP estimates and a specific MCMC scheme we developed in [30] for computing CM estimates.

2.2.1. A 2D Deblurring Example We start with simple 2D image deblurring using the ℓ_1 prior, i.e., $p_{pr}(u) \propto \exp(-\lambda|u|_1)$. The unknown intensity function $\tilde{u} : [0, 1]^2 \rightarrow \mathbb{R}_+$ consists of a few of circular spots of constant intensity whose radii and intensities slightly vary between single spots (see Figure 5b). It is convoluted with a Gaussian kernel (standard deviation of 0.015), integrated over 1025×1025 regular pixels and contaminated by noise ($\sigma = 0.1 \cdot \|K\tilde{u}\|_\infty$). The resulting measurement data is displayed in Figure 5c. The image will be reconstructed using the ℓ_1 prior, i.e., $p_{pr}(u) \propto \exp(-\lambda|u|_1)$ on the same pixel grid used for the measurement. To avoid an inverse crime [21], the grid used for the generation of the measurement data was four times finer. The results are shown in Figure 6. While the MAP estimate is very close to \tilde{u} , the CM estimate is blurred and is not able to separate all intensity spots.

2.2.2. The Discretization Dilemma of the TV Prior As a second example, we choose a basic 1D scenario to examine *edge-preserving* image reconstruction using the TV prior, i.e., $p_{pr}(u) \propto \exp\left(-\lambda_n \sum_{i=1}^{n-1} |u_{i+1} - u_i|\right)$, where $u_i := u(t_i)$. We indexed

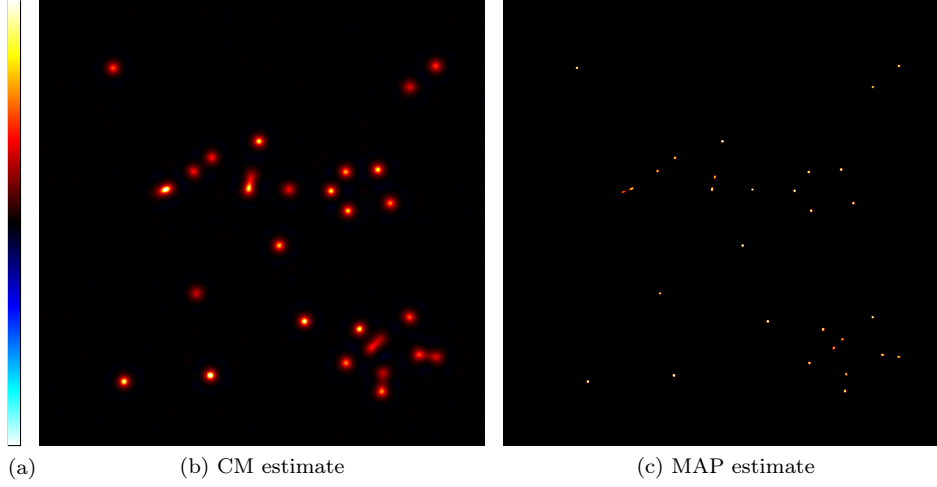


Figure 6: CM and MAP estimate for the 2D deblurring example.

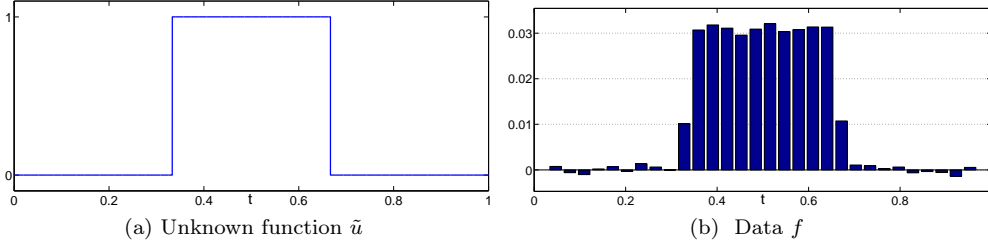


Figure 7: A simple 1D edge-preserving image reconstruction scenario.

λ_n by n to stress that we will choose it depending on the discretization level n . The unknown (light intensity) function $\tilde{u} : [0, 1] \rightarrow \mathbb{R}_+$ is the indicator function of $[\frac{1}{3}, \frac{2}{3}]$, see Figure 7a. It is integrated into m equidistant intervals (ccd pixels) and contaminated with noise (see Figure 7b). Further details can be found in [30].

This is a toy model for imaging applications where the task is to reconstruct an intensity image that is known to consist of piecewise homogeneous parts with sharp edges(cf. [4]). In Bayesian inversion, the use of TV priors stimulated interesting developments: In [27] it was shown that it is not possible to formulate the TV prior in a *discretization invariant* way, i.e., such that the posterior converges to a well defined limit probability density when n is increased while still reflecting the a priori information of edge-preservation. To summarize their results:

- For $n \rightarrow \infty$ the posterior only converges for $\lambda_n \propto \sqrt{n}$. However, its limit is a Gaussian smoothness prior and the CM estimate converges to a smooth limit while the MAP estimate converges to constant function. This is illustrated in Figure 8, where we computed CM and MAP estimates up to $n = 2^{16} - 1$.
- For $\lambda_n = \text{const.}$ and $n \rightarrow \infty$ both posterior and CM estimate diverge, while the MAP estimate converges to an edge-preserving limit, see Figure 9. Figure 10a shows a zoom to clarify the divergence of the CM estimate while Figure 10b

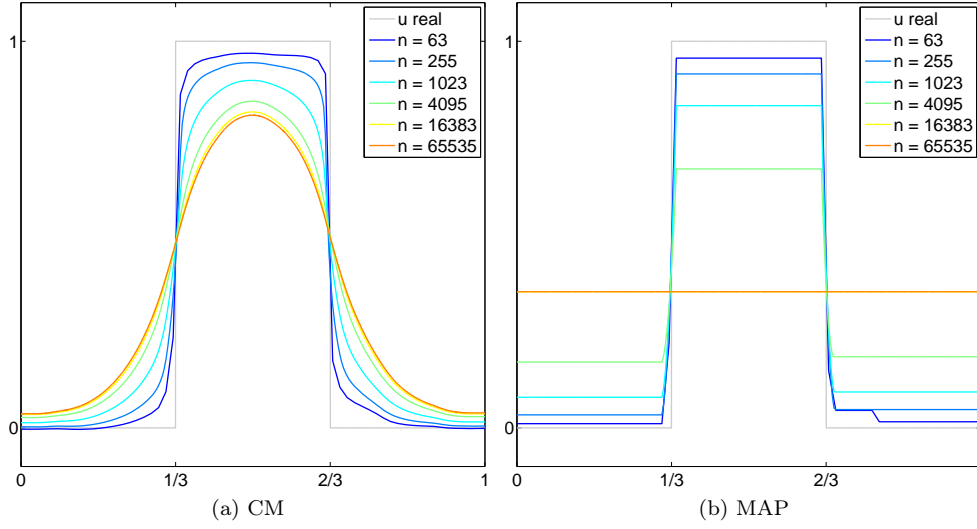


Figure 8: CM and MAP estimates (colored lines) for $\lambda_n \propto \sqrt{n+1}$ and increasing values of n vs. real solution $\tilde{u}(t)$ (gray line)

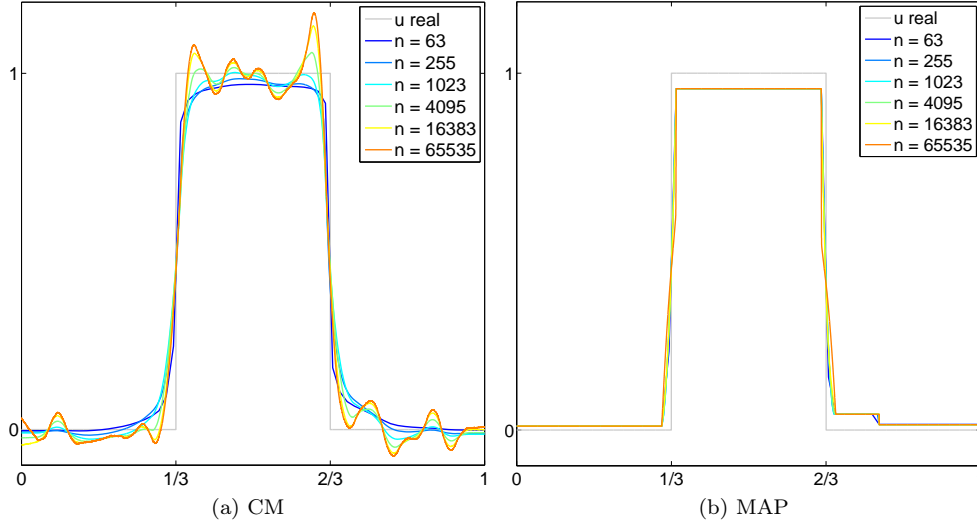


Figure 9: CM and MAP estimates (colored lines) for $\lambda_n = \text{const.}$ and increasing values of n vs. real solution $\tilde{u}(t)$ (gray line)

demonstrates that this is not an error of the MCMC sampler to compute it by comparing two CM estimates computed from independent MCMC chains.

2.2.3. Limited Angle CT with Besov Priors The discretization dilemma of the TV prior led to a search for edge-preserving and discretization invariant priors. In [26, 24, 16] *Besov space priors*, which rely on a weighted ℓ_1 norm of wavelet

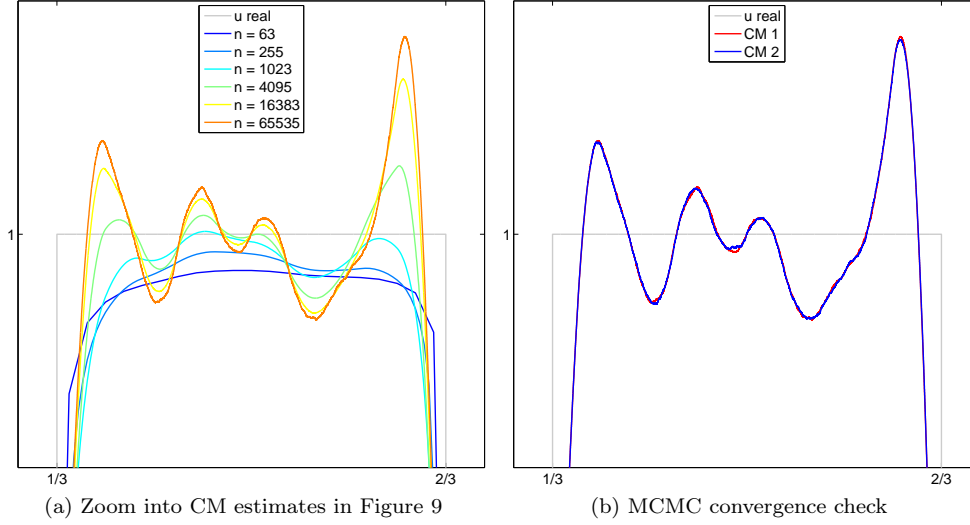


Figure 10: Details of CM estimates in Figure 9.

basis coefficients were shown to have these properties. Similar to [16], we will use Besov priors to reconstruct the Shepp-Logan phantom image \tilde{u} (see Figure 11a) from the integration of its Radon-projections into 500 noisy measurement pixel ($\sigma = 0.01 \cdot \|K\tilde{u}\|_\infty$) using only 45 projection angles (see Figure 11b). In Figure 12, CM and MAP estimates for increasing n are shown. As in [24], a Haar wavelet base was used to implement the Besov prior and λ was chosen by the *S-curve* method, i.e., depending on the sparsity of the true solution. While a closer examination of such scenarios with real data is subject to future research, for this article, the comparison between CM and MAP estimates is most important: In line with [26, 24, 16], we can confirm that both estimates converge in the limit of $n \rightarrow \infty$ and, more importantly, that they are nearly identical.

2.2.4. Summary of Observations and Discussions

- For Gaussian priors, MAP and CM estimates coincide.
- For the first two ℓ_1 -type priors we examined the MAP estimates were more convincing. For the Besov prior, MAP and CM estimate visually coincide. These findings are similar to a large variety of computational experiments. One could exaggerate that if a CM estimate looks good, it looks like the MAP estimate.
- In [31], we used hierarchical Bayesian modeling (cf. Section 1.2) for EEG source imaging. While the multimodality of the posterior complicates inference for such priors, suitably computed MAP estimates, again, outperformed the CM estimates.
- Recently, [12, 14, 29] revealed that every CM estimate for a prior $p(u)$ is also a MAP estimate for a different prior $\tilde{p}(u)$. Their intention was to warn against the common "reverse reading" of designing a particular $\mathcal{J}(u)$ for recovering certain classes of real solutions with (8) and then claiming to perform Bayesian MAP estimation with the prior $p(u) = \exp(-\lambda\mathcal{J}(u))$. In [13], it was shown that this MAP estimate is usually not very well suited to recover solutions u that are

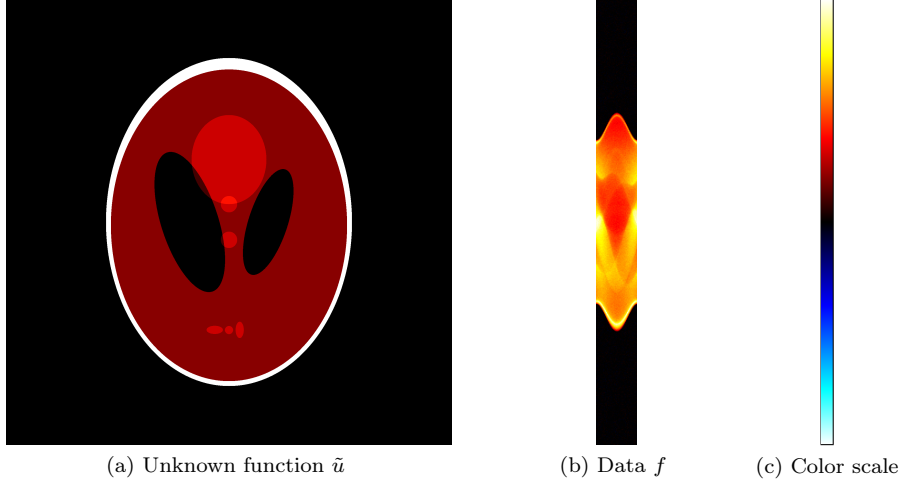


Figure 11: A 2D limited angle computerized tomography imaging scenario.

really distributed like $\exp(-\lambda\mathcal{J}(u))$ (cf. Figure 2, none of the true solutions used in the computational scenarios fitted to the assumed priors!). For the discussion in this article, these results mean that a general discrimination of MAP estimates based on the Bayes cost formalism only makes sense if one strongly believes that the chosen prior most accurately models the distribution of the real solution. Otherwise, one ends up in the contradiction that the appraised CM estimate will simultaneously be a discredited MAP estimate (just for another prior).

The next sections will present new theoretical ideas that resolve the contradictions between these observations and the classical view on the comparison between MAP and CM estimates.

3. A Novel Characterization of the MAP Estimate

This section presents a novel Bayes cost approach to MAP estimates for log-concave priors of the form $\exp(-\lambda\mathcal{J}(u))$. Throughout this section we assume that $\mathcal{J} : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ is a Lipschitz-continuous convex functional, such that for $\lambda > 0$ the function $u \mapsto \|Ku\|^2 + \lambda\mathcal{J}(u)$ has at least linear growth at infinity. Due to Rademacher's theorem (cf. [10]) this implies that $p_{po}(u|f)$ is log-concave and differentiable almost everywhere in \mathbb{R}^n . The main ingredient will be the (generalized) *Bregman distance*:

Definition 1. For a convex functional $\mathcal{J} : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$, the Bregman distance $D_{\mathcal{J}}^q(u, v)$ between $u, v \in \mathbb{R}^n$ for a subgradient $q \in \partial\mathcal{J}(v)$ is defined as

$$D_{\mathcal{J}}^q(u, v) = \mathcal{J}(u) - \mathcal{J}(v) - \langle q, u - v \rangle, \quad q \in \partial\mathcal{J}(v) \quad (15)$$

Note that if $\mathcal{J}(u)$ is Fréchet-differentiable in v , q is the Fréchet derivative $\mathcal{J}'(v)$. We will therefore simplify the notation to $D_{\mathcal{J}}(u, v)$, and use $D_{\mathcal{J}}^q(u, v)$ only if we want to stress the potential ambiguity. Table 1 lists the Bregman distances induced by some Gibbs energies $\mathcal{J}(u)$. Figure 13 gives an illustration: Basically, $D_{\mathcal{J}}(u, v)$ measures the difference between \mathcal{J} and its linearization in u at another point v . Further, $D_{\mathcal{J}}(u, v) \geq 0$ and for strictly convex $\mathcal{J}(u)$, $D_{\mathcal{J}}(u, v) = 0$ implies $u = v$. However, the

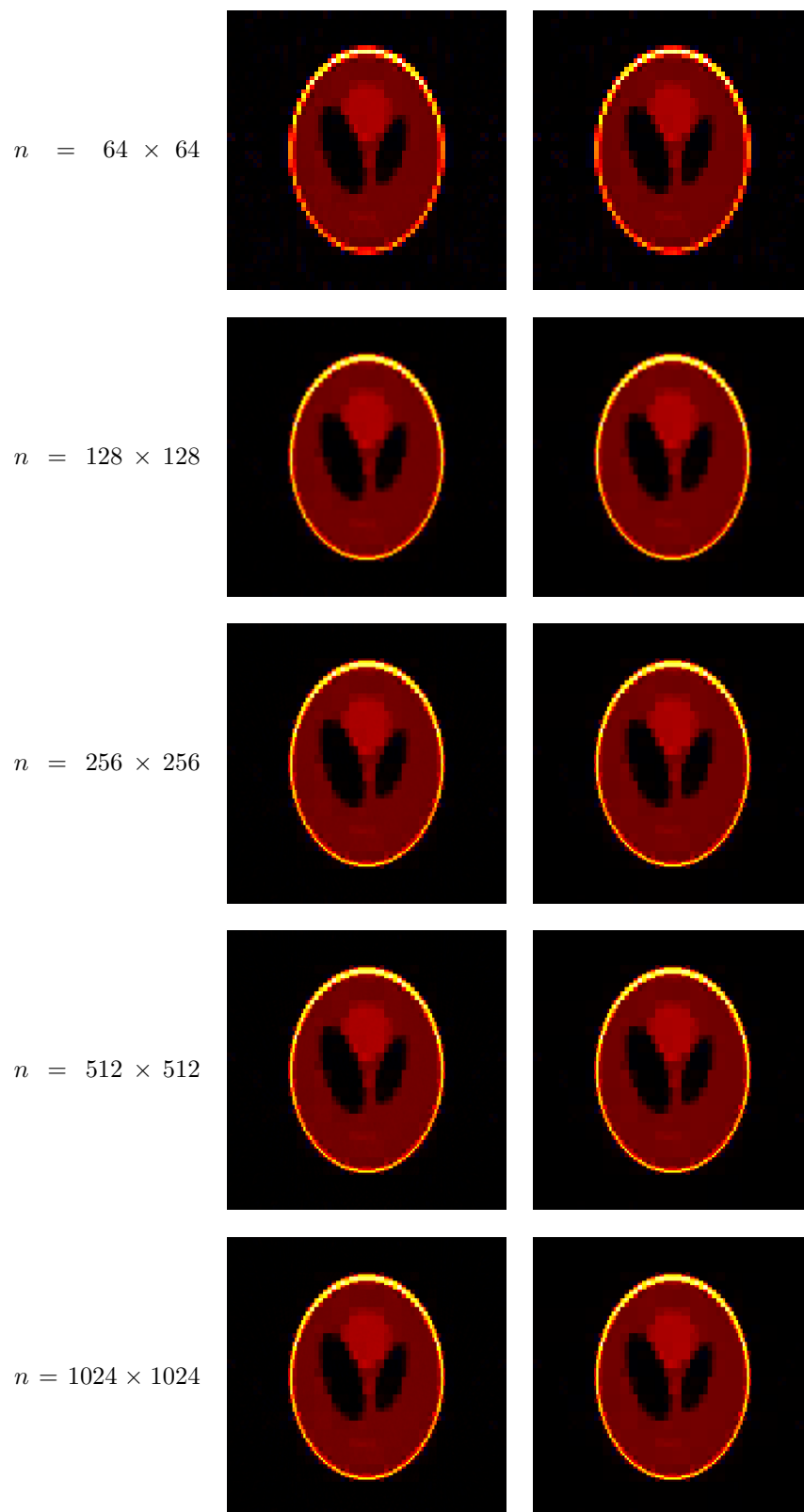


Figure 12: CM (left image column) and MAP (right image column) estimates for increasing resolution of the reconstruction grid.

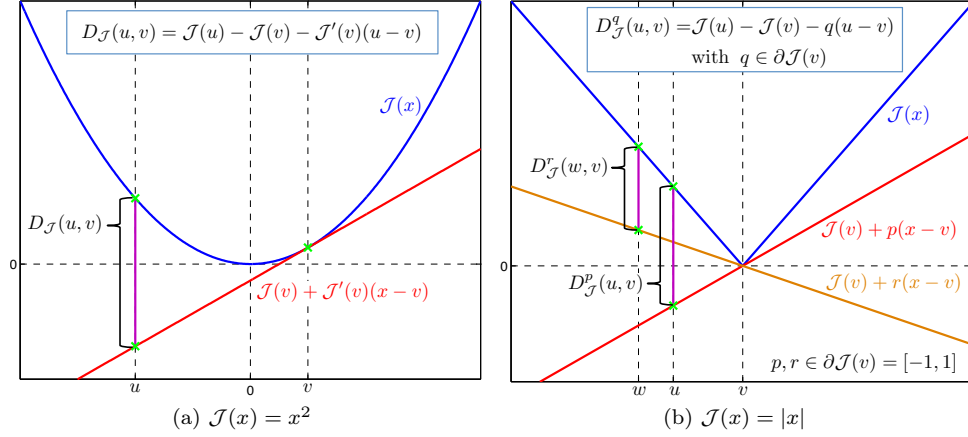


Figure 13: Illustrative explanation of the Bregman distance.

 Table 1: Bregman distances induced by some Gibbs energies $\mathcal{J}(u)$ commonly used for prior modeling. Note that if \mathcal{J} is separable, so is $D_{\mathcal{J}}^q(u, v)$. In these cases, the scalar expression are listed, only.

$\mathcal{J}(u)$	$\text{dom}(\mathcal{J})$	$D_{\mathcal{J}}(u, v)$
$\frac{1}{2}\ Lu\ _2^2$	\mathbb{R}^n	$\frac{1}{2}\ L(u-v)\ _2^2$
$ u ^p, (1 < p < \infty)$	\mathbb{R}	$ u ^p - p u \text{sign}(v) v ^{p-1} + (p-1) v ^p$
$ u $	\mathbb{R}	$(\text{sign}(u) - \text{sign}(v))u$
$u \log u - u$	$\mathbb{R}_{\geq 0}$	$u \log \frac{u}{v} + v - u$ (Kullback-Leibler divergence)

Bregman distance is not a distance in the usual mathematical sense, i.e., a metric, as it is, in general, neither symmetric nor satisfies the triangle inequality. We will further use that $D_{\mathcal{J}}(u, v)$ is convex in u . Bregman distances have become an important tool in variational regularization, e.g., to derive error estimates and convergence rates [3, 2], to enhance inverse methods by Bregman iterations [5, 32] or to develop optimization schemes like the Split-Bregman algorithm [11] used in this paper.

3.1. New Bayes Cost Functions

The classical discrimination of the MAP estimate as only being asymptotically a Bayes estimator for the uniform cost (14) (cf. Section 2.1) has a crucial flaw: It does not mean that the MAP estimate cannot be a proper Bayes estimator for a different cost function. This motivates to look for alternative costs better suited to the asymptotic Banach space structure such as Bregman distance costs:

Definition 2. Let $L \in \mathbb{R}^{n \times n}$ be regular and $\beta > 0$. Define

$$\Psi_{LS}(u, \hat{u}) := \|K(\hat{u} - u)\|_{\Sigma_{\varepsilon}^{-1}}^2 + \beta \|L(\hat{u} - u)\|_2^2 \quad (16)$$

$$\Psi_{Brg}(u, \hat{u}) := \|K(\hat{u} - u)\|_{\Sigma_{\varepsilon}^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}, u) \quad (17)$$

Both $\Psi_{LS}(u, \hat{u})$ and $\Psi_{Brg}(u, \hat{u})$ are proper, convex (with respect to \hat{u}) cost

functions. In the following, we will need the decay property

$$\lim_{R \rightarrow \infty} \int_{\partial \mathcal{B}_R(0)} p_{po}(u|f) \, du = 0 \quad (18)$$

which is fulfilled under the linear growth assumption above, which yields for constants A, B independent of R

$$p_{po}(u|f) \leq A e^{-\frac{B}{R}} \quad \text{on } \mathcal{B}_R(0).$$

Theorem 1. *Let \mathcal{J} be as above and let $\lambda > 0$ and $\beta \geq 0$. Then the CM estimate is a Bayes estimator for $\Psi_{LS}(u, \hat{u})$ and the MAP estimate is a Bayes estimator for $\Psi_{Brg}(u, \hat{u})$.*

Proof. We start from (12) and insert the definition of $\Psi_{LS}(u, \hat{u})$:

$$\hat{u}_{\Psi_{LS}}(f) = \operatorname{argmin}_{\hat{u}} \left\{ \int \left(\|K(\hat{u} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + \beta \|L(\hat{u} - u)\|_2^2 \right) p_{po}(u|f) \, du \right\}$$

We can rewrite the above by inserting \hat{u}_{CM} and expanding squares

$$\begin{aligned} \hat{u}_{\Psi_{LS}}(f) = \operatorname{argmin}_{\hat{u}} \left\{ \int \left(\|K(\hat{u} - \hat{u}_{CM})\|_{\Sigma_\varepsilon^{-1}}^2 + \beta \|L(\hat{u} - \hat{u}_{CM})\|_2^2 \right) p_{po}(u|f) \, du \right. \\ \left. + \int \left(\|K(u - \hat{u}_{CM})\|_{\Sigma_\varepsilon^{-1}}^2 + \beta \|L(u - \hat{u}_{CM})\|_2^2 \right) p_{po}(u|f) \, du \right. \\ \left. - 2 \int \left(\langle K(\hat{u} - \hat{u}_{CM}), K(u - \hat{u}_{CM}) \rangle_{\Sigma_\varepsilon^{-1}} + \beta \langle L(\hat{u} - \hat{u}_{CM}), L(u - \hat{u}_{CM}) \rangle_2 \right) p_{po}(u|f) \, du \right\} \end{aligned}$$

Due to the linearity and the definition of the CM estimate (9) the last integral vanishes and hence, $\hat{u} = \hat{u}_{CM}$ is obviously a minimizer. For the MAP estimate, we again start from (12) and insert the definition of $\Psi_{Brg}(u, \hat{u})$:

$$\hat{u}_{\Psi_{Brg}}(f) = \operatorname{argmin}_{\hat{u}} \left\{ \int_{\mathbb{R}^n} \left(\|K(\hat{u} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}, u) \right) p_{po}(u|f) \, du \right\}$$

Now, we can exclude the null-set where $\mathcal{J}(u)$ is not Fréchet-differentiable,

$$\mathcal{S} := \{u \in \mathbb{R}^n \mid |\partial \mathcal{J}(u)| \neq 1\},$$

from the integration and insert the definition of $D_{\mathcal{J}}(\hat{u}, u)$ on \mathcal{S}^c :

$$\hat{u}_{\Psi_{Brg}}(f) = \operatorname{argmin}_{\hat{u}} \left\{ \int_{\mathcal{S}^c} \left(\|K(\hat{u} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda (\mathcal{J}(\hat{u}) - \mathcal{J}(u) - \langle \mathcal{J}'(u), \hat{u} - u \rangle) \right) p_{po}(u|f) \, du \right\}$$

The squared norm can be developed as in the case of the CM-estimate, while for the Bregman distance we use the following elementary identity:

$$D_{\mathcal{J}}(\hat{u}, u) = D_{\mathcal{J}}(\hat{u}, \hat{u}_{MAP}) + D_{\mathcal{J}}(\hat{u}_{MAP}, u) + \langle \hat{p}_{MAP} - \mathcal{J}'(u), \hat{u} - \hat{u}_{MAP} \rangle$$

Thus, on \mathcal{S}^c we have

$$\begin{aligned} & \|K(\hat{u} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}, u) \\ &= \|K(\hat{u} - \hat{u}_{MAP})\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}, \hat{u}_{MAP}) + \|K(\hat{u}_{MAP} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}_{MAP}, u) \\ & \quad + 2\langle K(\hat{u} - \hat{u}_{MAP}), K(\hat{u}_{MAP} - u) \rangle_{\Sigma_\varepsilon^{-1}} + 2\lambda \langle \hat{p}_{MAP} - \mathcal{J}'(u), \hat{u} - \hat{u}_{MAP} \rangle. \end{aligned}$$

The first two terms in the second line are obviously minimal for $\hat{u} = \hat{u}_{\text{MAP}}$, while the other terms in this line are independent of \hat{u} . In the last line we can insert the subgradient from the optimality condition for \hat{u}_{MAP} ,

$$\hat{p}_{\text{MAP}} = -\frac{1}{\lambda} K^* \Sigma_\varepsilon^{-1} (K \hat{u}_{\text{MAP}} - f) \in \partial \mathcal{J}(\hat{u}_{\text{MAP}}),$$

and rewrite

$$\begin{aligned} 2 \langle K(\hat{u} - \hat{u}_{\text{MAP}}), K(\hat{u}_{\text{MAP}} - u) \rangle_{\Sigma_\varepsilon^{-1}} + 2\lambda \langle \hat{p}_{\text{MAP}} - \mathcal{J}'(u), \hat{u} - \hat{u}_{\text{MAP}} \rangle \\ = -2 \langle K(\hat{u} - \hat{u}_{\text{MAP}}), Ku - f \rangle_{\Sigma_\varepsilon^{-1}} - 2\lambda \langle -\mathcal{J}'(u), \hat{u} - \hat{u}_{\text{MAP}} \rangle \\ = -2 \langle K^* \Sigma_\varepsilon^{-1} (Ku - f) + \lambda \mathcal{J}'(u), \hat{u} - \hat{u}_{\text{MAP}} \rangle \\ = 2 \langle \nabla_u \log p_{po}(u|f), \hat{u} - \hat{u}_{\text{MAP}} \rangle. \end{aligned}$$

Using the logarithmic derivative $\nabla_u p_{po}(u|f) = (\nabla_u \log p_{po}(u|f)) p_{po}(u|f)$, the posterior expectation of the latter equals

$$2 \int_{S^c} \langle \nabla_u \log p_{po}(u|f), \hat{u} - \hat{u}_{\text{MAP}} \rangle p_{po}(u|f) du = 2 \langle \int_{S^c} \nabla_u p_{po}(u|f) du, \hat{u} - \hat{u}_{\text{MAP}} \rangle.$$

With Gauss' theorem and (18) we finally obtain:

$$\begin{aligned} \left\| \int \nabla_u p_{po}(u|f) du \right\| &= \lim_{R \rightarrow \infty} \left\| \int_{\mathcal{B}_R(0)} \nabla_u p_{po}(u|f) du \right\| \\ &= \lim_{R \rightarrow \infty} \left\| \int_{\partial \mathcal{B}_R(0)} p_{po}(u|f) \frac{u}{R} du \right\| \\ &\leq \lim_{R \rightarrow \infty} \int_{\partial \mathcal{B}_R(0)} p_{po}(u|f) du \\ &= 0 \end{aligned}$$

□

First, we apply Theorem 1 to the fundamental case of Gaussian priors. We can parameterize any (centered) Gaussian energy as $\mathcal{J}(u) = \beta/(2\lambda) \|Lu\|_2^2$. For this choice $2\lambda D_{\mathcal{J}}(\hat{u}, u) = \beta \|L(\hat{u} - u)\|_2^2$, and $\Psi_{\text{LS}}(u, \hat{u}) = \Psi_{\text{Brg}}(u, \hat{u})$: The equality of MAP and CM estimate in the Gaussian case is no longer a strange coincidence but follows naturally from the properties of the Bregman distance.

In the non-Gaussian case, the domain of \mathcal{J} usually defines a Banach space or a subset thereof in the limit $n \rightarrow \infty$. E.g., the discrete total variation prior will define the space of functions of bounded variation in the limit. In such a space there is no natural Hilbert space norm that one should obtain as the limit of $\|Lu\|^2$. Even worse, it is questionable whether any Hilbert space norm is a meaningful measure for functions of bounded variation. The only reasonable choice might be $L = 0$, which means that $\Psi_{\text{LS}}(u, \hat{u})$ measures purely in the output space, which will be a Hilbert space. However, for ill-posed inverse problems with noisy data it is well-established that one should not just minimize a criterion related to the output Ku .

3.2. A MAP-Centered Form of the Posterior

As pointed out in Section 2.1, one classical geometrical argument was that the CM estimate is in the center of mass of $p_{po}(u|f)$ while the MAP estimate does not allow for such an interpretation (cf. Section 2.1). Using Bregman distances, we can rewrite the $p_{po}(u|f)$ in a MAP-centered form, which also disqualifies this argument. We use the optimality condition of the MAP-estimate (8),

$$K^* \Sigma_\varepsilon^{-1} (K \hat{u}_{\text{MAP}} - f) + \lambda \hat{p}_{\text{MAP}} = 0, \quad \hat{p}_{\text{MAP}} \in \partial \mathcal{J}(\hat{u}_{\text{MAP}}), \quad (19)$$

to rewrite $K^* \Sigma_\varepsilon^{-1} f$ in the posterior energy:

$$\begin{aligned} & \frac{1}{2} \|K u - f\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda \mathcal{J}(u) \\ &= \frac{1}{2} \|K u\|_{\Sigma_\varepsilon^{-1}}^2 - \langle K^* \Sigma_\varepsilon^{-1} f, u \rangle + \lambda \mathcal{J}(u) + \frac{1}{2} \|f\|_{\Sigma_\varepsilon^{-1}}^2 \\ &= \frac{1}{2} \|K u\|_{\Sigma_\varepsilon^{-1}}^2 - \langle K^* \Sigma_\varepsilon^{-1} K \hat{u}_{\text{MAP}} + \lambda \hat{p}_{\text{MAP}}, u \rangle + \lambda \mathcal{J}(u) + \frac{1}{2} \|f\|_{\Sigma_\varepsilon^{-1}}^2 \\ &= \frac{1}{2} \|K u\|_{\Sigma_\varepsilon^{-1}}^2 - \langle \Sigma_\varepsilon^{-1} K \hat{u}_{\text{MAP}}, K u \rangle + \frac{1}{2} \|K \hat{u}_{\text{MAP}}\|_{\Sigma_\varepsilon^{-1}}^2 \\ &\quad + \lambda (\mathcal{J}(u) - \mathcal{J}(\hat{u}_{\text{MAP}}) - \langle \hat{p}_{\text{MAP}}, u - \hat{u}_{\text{MAP}} \rangle) \\ &\quad - \frac{1}{2} \|K \hat{u}_{\text{MAP}}\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda (\mathcal{J}(\hat{u}_{\text{MAP}}) - \langle \hat{p}_{\text{MAP}}, \hat{u}_{\text{MAP}} \rangle) + \frac{1}{2} \|f\|_{\Sigma_\varepsilon^{-1}}^2 \\ &= \frac{1}{2} \|K(u - \hat{u}_{\text{MAP}})\|_{\Sigma_\varepsilon^{-1}}^2 + \lambda D_{\mathcal{J}}^{\hat{p}_{\text{MAP}}}(u, \hat{u}_{\text{MAP}}) + \text{const.}, \end{aligned} \quad (20)$$

where const. sums all terms not depending on u . Hence, we can write the posterior as

$$p_{po}(u|f) \propto \exp \left(-\frac{1}{2} \|K(u - \hat{u}_{\text{MAP}})\|_{\Sigma_\varepsilon^{-1}}^2 - \lambda D_{\mathcal{J}}^{\hat{p}_{\text{MAP}}}(u, \hat{u}_{\text{MAP}}) \right). \quad (21)$$

Now, the posterior energy is sum of two convex functionals both minimized by \hat{u}_{MAP} , i.e., \hat{u}_{MAP} is the center of $p_{po}(u|f)$ with respect to the distance induced by (20).

3.3. Average Optimality of the CM Estimate

To further compare MAP and CM estimates, we derive an ‘‘average optimality condition’’ for the CM estimate. Let

$$\hat{p}_{\text{CM}} := \mathbb{E}[\mathcal{J}'(u)] = \int \mathcal{J}'(u) p_{po}(u|f) du \quad (22)$$

be the CM estimate for the (sub)gradient of $\mathcal{J}(u)$. We have:

$$\begin{aligned} K^* \Sigma_\varepsilon^{-1} (K \hat{u}_{\text{CM}} - f) + \lambda \hat{p}_{\text{CM}} &= K^* (K \Sigma_\varepsilon^{-1} \mathbb{E}[u] - f) + \lambda \mathbb{E}[\mathcal{J}'(u)] \\ &= \mathbb{E} [K^* \Sigma_\varepsilon^{-1} (K u - f) + \lambda \mathcal{J}'(u)] \\ &= \int_{\mathcal{S}^c} K^* \Sigma_\varepsilon^{-1} (K u - f) + \lambda \mathcal{J}'(u) p_{po}(u|f) du \\ &= \int_{\mathcal{S}^c} \nabla_u p_{po}(u|f) du = 0, \end{aligned} \quad (23)$$

where the integral term, again, vanishes. Comparing (23) to (19) we see that the CM estimate fulfills an optimality condition ‘‘on average’’, i.e., with respect to the average gradient $\hat{p}_{\text{CM}} = \mathbb{E}[\mathcal{J}'(u)]$ but not with respect to the gradient $\mathcal{J}'(\hat{u}_{\text{CM}}) = \mathcal{J}'(\mathbb{E}[u])$. The difference between MAP and CM estimate here manifests in $\mathcal{J}'(\mathbb{E}[u]) \neq \mathbb{E}[\mathcal{J}'(u)]$, which, again, vanishes for the Gaussian case where $\mathcal{J}'(u)$ is linear.

3.4. New Inequalities

Finally, we show that when measured in the Bregman distance $D_J(\hat{u}, u)$, which is a more reasonable error measure than norms in the case of a non-quadratic $\mathcal{J}(u)$, the MAP estimate performs better than the CM estimate. In return, the CM estimate out-performs the MAP estimate when the error is measured in a quadratic distance:

Theorem 2. *Let $L \in \mathbb{R}^{n \times n}$ be regular, then we have*

$$\mathbb{E} [\|L(\hat{u}_{CM} - u)\|_2^2] \leq \mathbb{E} [\|L(\hat{u}_{MAP} - u)\|_2^2] \quad (24)$$

$$\mathbb{E} [D_{\mathcal{J}}(\hat{u}_{MAP}, u)] \leq \mathbb{E} [D_{\mathcal{J}}(\hat{u}_{CM}, u)] \quad (25)$$

Proof. The first inequality directly follows from the fact that \hat{u}_{CM} is also the Bayes estimator for $\Psi(u, \hat{u}) = \|L(\hat{u}_{CM} - u)\|_2^2$, which follows from the proof to Theorem 1. For the second inequality, we use the minimizing properties of MAP and CM estimates:

$$\begin{aligned} & \int \left(\|K(\hat{u}_{MAP} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}_{MAP}, u) \right) p_{po}(u|f) du \\ & \leq \int \left(\|K(\hat{u}_{CM} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}_{CM}, u) \right) p_{po}(u|f) du \\ & \leq \int \left(\|K(\hat{u}_{MAP} - u)\|_{\Sigma_\varepsilon^{-1}}^2 + 2\lambda D_{\mathcal{J}}(\hat{u}_{CM}, u) \right) p_{po}(u|f) du \\ & \quad + \beta \int (\|L(\hat{u}_{MAP} - u)\|_2^2 - \|L(\hat{u}_{CM} - u)\|_2^2) p_{po}(u|f) du \end{aligned}$$

Since $\beta > 0$ is arbitrary, we can consider $\beta \rightarrow 0$ and obtain

$$\int D_{\mathcal{J}}(\hat{u}_{MAP}, u) p_{po}(u|f) du \leq \int D_{\mathcal{J}}(\hat{u}_{CM}, u) p_{po}(u|f) du$$

□

4. Discussion and Conclusions

In this article, we examined point estimates in Bayesian inversion from both computational and theoretical perspectives and contrasted recent observations with classical assumptions. We showed that the common discrimination of MAP estimates based on the Bayes cost formalism is not valid: Using Bregman distances, the MAP estimate is a proper Bayes estimator for a convex cost function as well (Section 3.1). Further aspects like the centeredness of the posterior around the MAP estimate (Section 3.2) and the optimality with respect to error measures (Section 3.4) were examined as well.

A potential irritation might be that the cost function for the MAP estimate depends on the chosen prior while the one for the CM estimate does not. However, this is usually not a drawback but rather an advantage: $\mathcal{J}(u)$ is chosen such that it grasps the most distinctive features of u . Often, one is consequently also most interested in estimating these features correctly, which is measured by $D_{\mathcal{J}}(u, v)$ better than in some squared error metric. For instance, in a situation like the 2D image deblurring scenario in Section 2.2.1, one is mainly interested in the correct separation and location of the intensity spots while their absolute amplitudes might be of minor interest. In such situations, the standard squared error is a poor indicator of reconstruction quality (see also the discussions in [2, 3, 5, 35]). On the other hand, the induced Bregman

distance $D_{\mathcal{J}}(u, v)$ is 0 if the sign pattern of u and v coincide (cf. Table 1) and grows only linearly, otherwise.

The main aim of this article was to rehabilitate the MAP estimate for Bayesian inversion and, thereby, to also disprove common misconceptions about the nature of MAP estimation. We think that the "MAP vs. CM" debate is *not* an interesting question for relating variational regularization and Bayesian inference. It might be a very obvious question but it puts the focus on a direct comparison between point estimates and suggests to choose between one of the two approaches. The real strength of Bayesian approaches is to model and quantify uncertainty and information at all stages of the problem, *beyond* point estimates. In this direction, Bayesian techniques can very well complement variational approaches.

In addition, Bregman distances have proven to be an interesting tool to analyze Bayesian inversion as well, which further emphasizes the strength of this concept for inverse problems theory.

A secondary objective of this article was to demonstrate that sample-based Bayesian inversion is feasible in high dimensions if suitable computational tools are available: The dimensions of u in the computed examples were intentionally chosen to be very high. Although only CM estimates were computed here, other Bayesian inference techniques can be implemented using posterior samples (which might also be more interesting, as the main conclusion suggests).

Going from discrete to infinite dimensional Bayesian inversion has recently attracted attention [26, 18, 17, 36, 7] for theoretical reasons as well as for designing algorithms that work in high dimensional settings. Extending the ideas presented here to infinite dimensions could also be useful to draw further connections to variational approaches, which are often rather formulated and analyzed in a function space setting.

This article investigated log-concave priors, which covers many priors used in Bayesian inversion. However, especially for implementing sparsity constraints priors that do not fit into this category are used more and more often (cf. Section 1.2). As their use might lead to multimodal posteriors, getting more insight into the relation of the CM estimate and the local maxima of the posterior would be very valuable.

Connected to the last point is the extension to non-linear inverse problems.

Finally, our presentation covered Gaussian noise, only. An extension to other relevant noise models like Poisson noise would be very interesting.

Acknowledgments

This work has been supported by the German Science Foundation (DFG) via grant BU 2327/6-1 within the *Inverse Problems Initiative* and via Cells in Motion Cluster of Excellence (EXC 1003 CiM), University of Münster. The authors thank Tapio Helin, Matti Lassas and Samuli Siltanen (all Helsinki University) for stimulating this research direction.

References

- [1] J. Bardsley, D. Calvetti, and E. Somersalo. Hierarchical regularization for edge-preserving reconstruction of PET images. *Inverse Probl.* 26:035010 (16pp), 2010.
- [2] M. Benning. *Singular Regularization of Inverse Problems*. PhD thesis, University of Muenster, 2011.
- [3] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Probl.* 20:1411–1421, 2004.

- [4] M. Burger and S. Osher. A Guide to the TV Zoo. In *Level Set and PDE Based Reconstruction Methods in Imaging*, Lecture Notes in Mathematics, pages 1–70. Springer International Publishing, 2013.
- [5] M. Burger, E. Resmerita, and L. He. Error estimation for bregman iterations and inverse scale space methods in image restoration. *Computing*, 81(2-3):109–135, 2007.
- [6] T. Cui, C. Fox, and M. J. O’Sullivan. Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. *Water Resour Res*, 47, 2011.
- [7] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. Map estimators and their consistency in bayesian nonparametric inverse problems. *Inverse Probl*, 29(9):095017, 2013.
- [8] T. Eltoft, T. Kim, and T.-W. Lee. On the multivariate Laplace distribution. *IEEE Signal Process Lett*, 13(5):300–303, may 2006.
- [9] H. Engl, M. Hanke-Bourgeois, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, Berlin, 1996.
- [10] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*, volume 5. CRC press, 1991.
- [11] T. Goldstein and S. Osher. The Split Bregman method for L1-regularized problems. *SIAM J Img Sci*, 2:323–343, April 2009.
- [12] R. Gribonval. Should Penalized Least Squares Regression be Interpreted as Maximum A Posteriori Estimation? *IEEE Trans Signal Process*, 59(5):2405–2410, 2011.
- [13] R. Gribonval, V. Cevher, and M. Davies. Compressible Distributions for High-Dimensional Statistics. *IEEE Trans Inf Theory*, 58(8):5016–5034, 2012.
- [14] R. Gribonval and P. Machart. Reconciling "priors" & "priors" without prejudice? Research Report RR-8366, INRIA, Sept. 2013.
- [15] H. Haario, M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen. Markov Chain Monte Carlo Methods for High Dimensional Inversion in Remote Sensing. *J R Stat Soc Series B Stat Methodol*, 66(3):591–607, 2004.
- [16] K. Hämmäläinen, A. Kallonen, V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen. Sparse Tomography. *SIAM J Sci Comput*, 35(3):B644–B665, 2013.
- [17] T. Helin. On infinite-dimensional hierarchical probability models in statistical inverse problems. *Inverse Probl Imaging*, 3:567–597, 2010.
- [18] T. Helin and M. Lassas. Hierarchical Models in Statistical Inverse Problems and the Mumford–Shah Functional. Technical Report arXiv:0908.3396, Aug 2009.
- [19] R. N. Henson, J. Mattout, C. Phillips, and K. J. Friston. Selecting forward models for MEG source-reconstruction using model-evidence. *Neuroimage*, 46(1):168–176, May 2009.
- [20] J. P. Kaipio and C. Fox. The Bayesian Framework for Inverse Problems in Heat Transfer. *Heat Transfer Eng.*, 32(9):718–753, 2011.
- [21] J. P. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer New York, 2005.
- [22] J. P. Kaipio and E. Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *J Comput Appl Math*, 198(2):493–504, 2007. Applied Computational Inverse Problems.
- [23] S. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Number Bd. 1 in Fundamentals of Statistical Signal Processing. Prentice-Hall PTR, 1998.
- [24] V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen. Sparsity-promoting Bayesian inversion. *Inverse Probl*, 28(2):025005 (28pp), 2012.
- [25] T. Lahivaara, A. Seppanen, J. Kaipio, J. Vauhkonen, L. Korhonen, T. Tokola, and M. Maltamo. Bayesian Approach to Tree Detection Based on Airborne Laser Scanning Data. *IEEE Trans Geosci Remote Sens*, PP(99):1–10, 2013.
- [26] M. Lassas, E. Saksman, and S. Siltanen. Discretization invariant Bayesian inversion and Besov space priors. *Inverse Probl Imaging*, (3):87–122, Jan 2009.
- [27] M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl*, 20:1537–1563, 2004.
- [28] A. Lipponen, V. Kolehmainen, S. Romakkaniemi, and H. Kokkola. Correction of approximation errors with random forests applied to modelling of cloud droplet formation. *Geosci Model Dev*, 6(6):2087–2098, 2013.
- [29] C. Louchet and L. Moisan. Posterior Expectation of the Total Variation model: Properties and Experiments. *SIAM J Imaging Sci*, 6(4):2640–2684, 2013.
- [30] F. Lucka. Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors. *Inverse Probl*, 28(12):125012, 2012.
- [31] F. Lucka, S. Pursiainen, M. Burger, and C. H. Wolters. Hierarchical Bayesian inference for

- the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents. *Neuroimage*, 61(4):1364–1382, 2012.
- [32] M. Moeller. *Multiscale Methods for Generalized Sparse Recovery and Applications in High Dimensional Imaging*. PhD thesis, University of Muenster, 2012.
- [33] A. Nissinen, V. Kolehmainen, and J. Kaipio. Compensation of Modelling Errors Due to Unknown Domain Boundary in Electrical Impedance Tomography. *IEEE Trans Med Imaging*, (99):1–1, 2011.
- [34] S. Pursiainen and M. Kaasalainen. Iterative alternating sequential (IAS) method for radio tomography of asteroids in 3D. *Planet Space Sci*, 82-83(0):84–98, 2013.
- [35] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. Kazimierski. *Regularization Methods in Banach Spaces*. Radon Series on Computational and Applied Mathematics. Walter De Gruyter Incorporated, 2012.
- [36] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 5 2010.
- [37] T. Tarvainen, A. Pulkkinen, B. Cox, J. Kaipio, and S. Arridge. Bayesian Image Reconstruction in Quantitative Photoacoustic Tomography. *IEEE Trans Med Imaging*, 32(12):2287–2298, Dec 2013.
- [38] U. Toussaint. Bayesian inference in physics. *Rev Mod Phys*, 83(3):943–999, 2011.
- [39] N. J. Trujillo-Barreto, E. Aubert-Vázquez, and P. A. Valdés-Sosa. Bayesian model averaging in EEG/MEG imaging. *Neuroimage*, 21(4):1300–1319, Apr. 2004.
- [40] Y. Wang, X. Jiang, B. Yu, and M. Jiang. A Hierarchical Bayesian Approach for Aerosol Retrieval Using MISR Data. *J Am Stat Assoc*, 108(502):483–493, 2013.